Permutational Rademacher Complexity A New Complexity Measure for Transductive Learning

Ilya Tolstikhin, Nikita Zhivotovskiy, Gilles Blanchard

Permutational Rademacher Complexity
A New Complexity Measure for Transductive Learning

Ilya Tolstikhin[1], Nikita Zhivotovskiy[23], and Gilles Blanchard[4]

[1] Max-Planck-Institute for Intelligent Systems, Tübingen, Germany
ilya@tuebingen.mpg.de
[2] Moscow Institute of Physics and Technology, Moscow, Russia
[3] Institute for Information Transmission Problems, Moscow, Russia
nikita.zhivotovskiy@phystech.edu
[4] Department of Mathematics, Universität Potsdam, Potsdam, Germany
gilles.blanchard@math.uni-potsdam.de

Abstract. Transductive learning considers situations when a learner observes m labelled training points and u unlabelled test points with the final goal of giving correct answers for the test points. This paper introduces a new complexity measure for transductive learning called *Permutational Rademacher Complexity* (PRC) and studies its properties. A novel symmetrization inequality is proved, which shows that PRC provides a tighter control over expected suprema of empirical processes compared to what happens in the standard i.i.d. setting. A number of comparison results are also provided, which show the relation between PRC and other popular complexity measures used in statistical learning theory, including Rademacher complexity and Transductive Rademacher Complexity (TRC). We argue that PRC is a more suitable complexity measure for transductive learning. Finally, these results are combined with a standard concentration argument to provide novel data-dependent risk bounds for transductive learning.

Keywords: Transductive Learning, Rademacher Complexity, Statistical Learning Theory, Empirical Processes, Concentration Inequalities

1 Introduction

Rademacher complexities ([14], [2]) play an important role in the widely used concentration-based approach to statistical learning theory [4], which is closely related to the analysis of empirical processes [21]. They measure a complexity of function classes and provide data-dependent risk bounds in the standard i.i.d. framework of inductive learning, thanks to symmetrization and concentration inequalities. Recently, a number of attempts were made to apply this machinery also to the *transductive learning* setting [22]. In particular, the authors of [10] introduced a notion of *transductive Rademacher complexity* and provided an extensive study of its properties, as well as general transductive risk bounds based on this new complexity measure.

In the transductive learning, a learner observes m labelled training points and u unlabelled test points. The goal is to give correct answers on the test points. Transductive learning naturally appears in many modern large-scale applications, including text mining, recommender systems, and computer vision, where often the objects to be classified are available beforehand. There are two different settings of transductive learning, defined by V. Vapnik in his book [22, Chap. 8]. The first one assumes that all the objects from the training and test sets are generated i.i.d. from an unknown distribution P. The second one is *distribution free*, and it assumes that the training and test sets are realized by a uniform and random partition of a fixed and finite general population of cardinality $N := m + u$ into two disjoint subsets of cardinalities m and u; moreover, no assumptions are made regarding the underlying source of this general population. The second setting has gained much attention[5] ([22], [9], [7], [10], [8], and [20]), probably due to the fact that any upper risk bound for this setting directly implies a risk bound also for the first setting [22, Theorem 8.1]. In essence, the second setting studies uniform deviations of risks computed on two disjoint finite samples. Following Vapnik's discussion in [6, p. 458], we would also like to emphasize that the second setting of transductive learning naturally appears as a middle step in proofs of the standard inductive risk bounds, as a result of symmetrization or the so-called *double-sample* trick. This way better transductive risk bounds also translate into better inductive ones.

An important difference between the two settings discussed above lies in the fact that the m elements of the training set in the second setting are interdependent, because they are sampled uniformly *without replacement* from the general population. As a result, the standard techniques developed for inductive learning, including concentration and Rademacher complexities mentioned in the beginning, can not be applied in this setting, since they are heavily based on the i.i.d. assumption. Therefore, it is important to study empirical processes in the setting of sampling without replacement.

Previous work. A large step in this direction was made in [10], where the authors presented a version of McDiarmid's bounded difference inequality [5] for sampling without replacement together with the Transductive Rademacher Complexity (TRC). As a main application the authors derived an upper bound on the binary test error of a transductive learning algorithm in terms of TRC. However, the analysis of [10] has a number of shortcomings. Most importantly, TRC depends on the unknown labels of the test set. In order to obtain computable risk bounds, the authors resorted to the contraction inequality [15], which is known to be a loose step [17], since it destroys any dependence on the labels.

Another line of work was presented in [20], where variants of Talagrand's concentration inequality were derived for the setting of sampling without replacement. These inequalities were then applied to achieve transductive risk bounds with fast rates of convergence $o(m^{-1/2})$, following a *localized* approach [1]. In contrast, in this work we consider only the worst-case analysis based on the

[5] For the extensive overview of transductive risk bounds we refer the reader to [18].

global complexity measures. An analysis under additional assumptions on the problem at hand, including Mammen-Tsybakov type low noise conditions [4], is an interesting open question and left for future work.

Summary of our results. This paper continues the analysis of empirical processes indexed by arbitrary classes of uniformly bounded functions in the setting of sampling without replacement, initiated by [10]. We introduce a new complexity measure called *permutational Rademacher complexity* (PRC) and argue that it captures the nature of this setting very well. Due to space limitations we present the analysis of PRC only for the special case when the training and test sets have the same size $m = u$, which is nonetheless sufficiently illustrative[6].

We prove a novel symmetrization inequality (Theorem 2), which shows that the expected PRC and the expected suprema of empirical processes when sampling without replacement are equivalent up to multiplicative constants. Quite remarkably, the new upper and lower bounds (the latter is often called *desymmetrization inequality*) both hold without any additive terms when $m = u$, in contrast to the standard i.i.d. setting, where an additive term of order $O(m^{-1/2})$ is unavoidable in the lower bound. For TRC even the upper symmetrization inequality [10, Lemma 4] includes an additive term of the order $O(m^{-1/2})$ and no desymmetrization inequality is known. This suggests that PRC may be a more suitable complexity measure for transductive learning. We would also like to note that the proof of our new symmetrization inequality is surprisingly simple, compared to the one presented in [10].

Next we compare PRC with other popular complexity measures used in statistical learning theory. In particular, we provide achievable upper and lower bounds, relating PRC to the conditional Rademacher complexity (Theorem 3). These bounds show that the PRC is upper and lower bounded by the conditional Rademacher complexity up to additive terms of orders $o(m^{-1/2})$ and $O(m^{-1/2})$ respectively, which are achievable (Lemma 1). In addition to this, Theorem 3 also significantly improves bounds on the complexity measure called *maximum discrepancy* presented in [2, Lemma 3]. We also provide a comparison between expected PRC and TRC (Corollary 1), which shows that their values are close up to small multiplicative constants and additive terms of order $O(m^{-1/2})$.

Finally, we apply these results to obtain a new computable data-dependent risk bound for transductive learning based on the PRC (Theorem 5), which holds for any bounded loss functions. We conclude by discussing the advantages of the new risk bound over the previously best known one of [10].

2 Notations

We will use calligraphic symbols to denote sets, with subscripts indicating their cardinalities: $\mathrm{card}(\mathcal{Z}_m) = m$. For any function f we will denote its average value computed on a finite set S by $\bar{f}(S)$. In what follows we will consider an arbitrary space $\mathcal{Z}$ (for instance, a space of input-output pairs) and class F of functions

[6] All the results presented in this paper are also available for the general $m \neq u$ case, but we defer them to a future extended version of this paper.

(for instance, loss functions) mapping $\mathcal{Z}$ to $\mathbb{R}$. Most of the proofs are deferred to the last section for improved readability.

Arguably, one of the most popular complexity measures used in statistical learning theory is the Rademacher complexity ([15], [14], [2]):

Definition 1 (Conditional Rademacher complexity). *Fix any subset $\mathcal{Z}_m = \{Z_1, \ldots, Z_m\} \subseteq \mathcal{Z}$. The following random quantity is commonly known as a* conditional Rademacher complexity*:*

$$\hat{R}_m(F, \mathcal{Z}_m) = \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\frac{2}{m} \sup_{f \in F} \sum_{i=1}^{m} \epsilon_i f(Z_i) \right],$$

where $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^m$ are i.i.d. Rademacher signs, taking values ± 1 with probabilities $1/2$. When the set $\mathcal{Z}_m$ is clear from the context we will simply write $\hat{R}_m(F)$.

As discussed in the introduction, Rademacher complexities play an important role in the analysis of empirical processes and statistical learning theory. However, this measure of complexity was devised mainly for the i.i.d. setting, which is different from our setting of sampling without replacement. The following complexity measure was introduced in [10] to overcome this issue:

Definition 2 (Transductive Rademacher complexity). *Fix any set $\mathcal{Z}_N = \{Z_1, \ldots, Z_N\} \subseteq \mathcal{Z}$, positive integers m, u such that $N = m + u$, and $p \in \left[0, \frac{1}{2}\right]$. The following quantity is called* Transductive Rademacher complexity *(TRC):*

$$\hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, p) = \left(\frac{1}{m} + \frac{1}{u} \right) \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in F} \sum_{i=1}^{N} \sigma_i f(Z_i) \right],$$

where $\boldsymbol{\sigma} = \{\sigma_1\}_{i=1}^{m+u}$ are i.i.d. random variables taking values ± 1 with probabilities p and 0 with probability $1 - 2p$.

We summarize the importance of these two complexity measures in the analysis of empirical processes when sampling without replacement in the following result:

Theorem 1. *Fix an N-element subset $\mathcal{Z}_N \subseteq \mathcal{Z}$ and let $m < N$ elements of $\mathcal{Z}_m$ be sampled uniformly without replacement from $\mathcal{Z}_N$. Also let m elements of $\mathcal{X}_m$ be sampled uniformly with replacement from $\mathcal{Z}_N$. Denote $\mathcal{Z}_u := \mathcal{Z}_N \setminus \mathcal{Z}_m$ with $u := \operatorname{card}(\mathcal{Z}_u) = N - m$. The following upper bound in terms of the i.i.d. Rademacher complexity was provided in [20]:*

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m} \sup_{f \in F} \left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m) \right) \leq \frac{N}{u} \cdot \mathop{\mathbb{E}}_{\mathcal{X}_m} \left[\hat{R}_m(F, \mathcal{X}_m) \right]. \tag{1}$$

The following bound in terms of TRC was provided in [10]. Assume that functions in F are uniformly bounded by B. Then for $p_0 := \frac{mu}{N^2}$ and $c_0 < 5.05$:

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m} \sup_{f \in F} \left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m) \right) \leq \hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, p_0) + c_0 B \frac{N \sqrt{\min(m, u)}}{mu}. \tag{2}$$

While (1) did not explicitly appear in [20], it can be immediately derived using [20, Corollary 8] and i.i.d. symmetrization of [13, Theorem 2.1].

Finally, we introduce our new complexity measure:

Definition 3 (Permutational Rademacher complexity). *Let $\mathcal{Z}_m \subseteq \mathcal{Z}$ be any fixed set of cardinality m. For any $n \in \{1, \dots, m-1\}$ the following quantity will be called a* permutational Rademacher complexity (PRC)*:*

$$\hat{Q}_{m,n}(F, \mathcal{Z}_m) = \mathop{\mathbb{E}}_{\mathcal{Z}_n} \sup_{f \in F} \left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n) \right),$$

where $\mathcal{Z}_n$ is a random subset of $\mathcal{Z}_m$ containing n elements sampled uniformly without replacement and $\mathcal{Z}_k := \mathcal{Z}_m \setminus \mathcal{Z}_n$. When the set $\mathcal{Z}_m$ is clear from the context we will simply write $\hat{Q}_{m,n}(F)$.

The name PRC is explained by the fact that if m is even then the definitions of $\hat{Q}_{m,m/2}(F)$ and $\hat{R}_m(F)$ are very similar. Indeed, the only difference is that the expectation in the PRC is over the randomly permuted sequence containing *equal number* of "-1" and "$+1$", whereas in Rademacher complexity the average is w.r.t. all the possible sequences of signs. The term "permutation complexity" has already appeared in [16], where it was used to denote a novel complexity measure for a model selection. However, this measure was specific to the i.i.d. setting and *binary* loss. Moreover, the bounds presented in [16] were of the same order as the risk bounds based on the Rademacher complexity with worse constants in the slack term.

3 Symmetrization and Comparison Results

We start with showing a version of the i.i.d. symmetrization inequality (references can be found in [15], [13]) for the setting of sampling without replacement. It shows that the expected supremum of empirical processes in this setting is up to multiplicative constants equivalent to the expected PRC.

Theorem 2. *Fix an N-element subset $\mathcal{Z}_N \subseteq \mathcal{Z}$ and let $m < N$ elements of $\mathcal{Z}_m$ be sampled uniformly without replacement from $\mathcal{Z}_N$. Denote $\mathcal{Z}_u := \mathcal{Z}_N \setminus \mathcal{Z}_m$ with $u := \operatorname{card}(\mathcal{Z}_u) = N - m$. If $m = u$ and m is even then for any $n \in \{1, \dots, m-1\}$:*

$$\frac{1}{2} \mathop{\mathbb{E}}_{\mathcal{Z}_m} \left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m) \right] \leq \mathop{\mathbb{E}}_{\mathcal{Z}_m} \sup_{f \in F} \left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m) \right) \leq \mathop{\mathbb{E}}_{\mathcal{Z}_m} \left[\hat{Q}_{m,n}(F, \mathcal{Z}_m) \right].$$

The inequalities also hold if we include absolute values inside the suprema.

Proof. The proof can be found in Sect. 5.1.

This inequality should be compared to the previously known complexity bounds of Theorem 1. First of all, in contrast to (1) and (2) the new bound provides a two sided control, which shows that PRC is a "correct" complexity measure for our setting. It is also remarkable that the lower bound (commonly known as

the *desymmetrization inequality*) does not include any additive terms, since in the standard i.i.d. setting the lower bound holds only up to an additive term of order $O(m^{-1/2})$ [13, Sect. 2.1]. Also note that this result does not assume the boundedness of functions in F, which is a necessary assumptions both in (2) and in the i.i.d. desymmetrization inequality.

Next we compare PRC with the conditional Rademacher complexity:

Theorem 3. *Let $\mathcal{Z}_m \subseteq \mathcal{Z}$ be any fixed set of even cardinality m. Then:*

$$\hat{Q}_{m,m/2}(F, \mathcal{Z}_m) \leq \left(1 + \frac{2}{\sqrt{2\pi m} - 2}\right) \hat{R}_m(F, \mathcal{Z}_m). \tag{3}$$

Moreover, if the functions in F are absolutely bounded by B then

$$\left|\hat{Q}_{m,m/2}(F, \mathcal{Z}_m) - \hat{R}_m(F, \mathcal{Z}_m)\right| \leq \frac{2B}{\sqrt{m}}. \tag{4}$$

The results also hold if we include absolute values inside suprema in $\hat{Q}_{m,n}, \hat{R}_m$.

Proof. Conceptually the proof is based on the coupling between a sequence $\{\epsilon_i\}_{i=1}^m$ of i.i.d. Rademacher signs and a uniform random permutation $\{\eta_i\}_{i=1}^m$ of a set containing $m/2$ plus and $m/2$ minus signs. This idea was inspired by the techniques used in [11]. The detailed proof can be found in Sect. 5.2.

Note that a typical order of $\hat{R}_m(F)$ is $O(m^{-1/2})$, thus the multiplicative upper bound (3) can be much tighter than the upper bound of (4). We would also like to note that Theorem 3 significantly improves bounds of Lemma 3 in [2], which relate the so-called *maximal discrepancy* measure of the class F to its Rademacher complexity (for the further discussion we refer to Appendix).

Our next result shows that bounds of Theorem 3 are essentially tight.

Lemma 1. *Let $\mathcal{Z}_m \subseteq \mathcal{Z}$ with even m. There are two finite classes F'_m and F''_m of functions mapping $\mathcal{Z}$ to $\mathbb{R}$ and absolutely bounded by 1, such that:*

$$\hat{Q}_{m,m/2}(F'_m, \mathcal{Z}_m) = 0, \quad (2m)^{-1/2} \leq \hat{R}_m(F'_m, \mathcal{Z}_m) \leq 2m^{-1/2}; \tag{5}$$

$$\hat{Q}_{m,m/2}(F''_m, \mathcal{Z}_m) = 1, \quad 1 - \sqrt{\frac{2}{\pi m}} \leq \hat{R}_m(F''_m, \mathcal{Z}_m) \leq 1 - \frac{4}{5}\sqrt{\frac{2}{\pi m}}. \tag{6}$$

Proof. The proof can be found in Sect. 5.3.

Inequalities (5) simultaneously show that (a) the order $O(m^{-1/2})$ of the additive bound (4) can not be improved, and (b) the multiplicative upper bound (3) can not be reversed. Moreover, it can be shown using (6) that the factor appearing in (3) can not be improved to $1 + o(m^{-1/2})$.

Finally, we compare PRC to the transductive Rademacher complexity:

Lemma 2. *Fix any set $\mathcal{Z}_N = \{Z_1, \ldots, Z_N\} \subseteq \mathcal{Z}$. If $m = u$ and $N = m + u$:*

$$\hat{R}_N(F, \mathcal{Z}_N) \leq \hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4) \leq 2\hat{R}_N(F, \mathcal{Z}_N).$$

Proof. The upper bound was presented in [10, Lemma 1]. For the lower bound, notice that if $p = 1/4$ the i.i.d. signs σ_i presented in Definition 2 have the same distribution as $\epsilon_i\eta_i$, where ϵ_i are i.i.d. Rademacher signs and η_i are i.i.d. Bernoulli random variables with parameters $1/2$. Thus, Jensen's inequality gives:

$$\hat{R}^{td}_{m+u}\left(F, \mathcal{Z}_N, 1/4\right) = \frac{4}{N} \mathop{\mathbb{E}}_{(\boldsymbol{\epsilon},\boldsymbol{\eta})} \left[\sup_{f\in F} \sum_{i=1}^{m+u} \epsilon_i\eta_i f(Z_i)\right] \geq \frac{4}{N} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{f\in F} \sum_{i=1}^{m+u} \epsilon_i \frac{1}{2} f(Z_i)\right].$$

Together with Theorems 2 and 3 this result shows that when $m = u$ the PRC can not be much larger than transductive Rademacher complexity:

Corollary 1. *Using notations of Theorem 2, we have:*

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \leq \left(2 + \frac{4}{\sqrt{2\pi N} - 2}\right) \hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4).$$

If functions in F are uniformly bounded by B then we also have a lower bound:

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \geq \frac{1}{2}\hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4) + \frac{2B}{\sqrt{N}}.$$

Proof. Simply notice that $\mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] = \hat{Q}_{N,m}(F, \mathcal{Z}_N)$.

4 Transductive Risk Bounds

Next we will use the results of Sect. 3 to obtain a new transductive risk bound. First we will shortly describe the setting.

We will consider the second, distribution-free setting of transductive learning described in the introduction. Fix any finite *general population* of input-output pairs $\mathcal{Z}_N = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are arbitrary input and output spaces. We make no assumptions regarding underlying source of $\mathcal{Z}_N$. The learner receives the labeled *training set* $\mathcal{Z}_m$ consisting of $m < N$ elements sampled uniformly without replacement from $\mathcal{Z}_N$. The remaining *test set* $\mathcal{Z}_u := \mathcal{Z}_N \setminus \mathcal{Z}_m$ is presented to the learner *without labels* (we will use $\mathcal{X}_u$ to denote the inputs of $\mathcal{Z}_u$). The goal of the learner is to find a predictor in the fixed *hypothesis class* $\mathcal{H}$ based on the training sample $\mathcal{Z}_m$ and unlabelled test points $\mathcal{X}_u$, which has a small test risk measured using bounded *loss function* $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [0,1]$. For $h \in \mathcal{H}$ and $(x, y) \in \mathcal{Z}_N$ denote $\ell_h(x, y) = \ell\big(h(x), y\big)$ and also denote the loss class $L_\mathcal{H} = \{\ell_h\colon h \in \mathcal{H}\}$. Then the test and training risks of $h \in \mathcal{H}$ are defined as $\mathrm{err}_u(h) := \overline{\ell_h}(\mathcal{Z}_u)$ and $\mathrm{err}_m(h) := \overline{\ell_h}(\mathcal{Z}_m)$ respectively.

Following risk bound in terms of TRC was presented in [10, Corollary 2]:

Theorem 4 ([10]). *If $m = u$ then with probability at least $1-\delta$ over the random training set $\mathcal{Z}_m$ any $h \in \mathcal{H}$ satisfies:*

$$\mathrm{err}_u(h) \leq \mathrm{err}_m(h) + \hat{R}^{td}_{m+u}\left(L_\mathcal{H}, \mathcal{Z}_N, 1/4\right) + 11\sqrt{\frac{2}{N}} + \sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \qquad (7)$$

Using results of Sect. 3 we obtain the following risk bound:

Theorem 5. *If $m = u$ and $n \in \{1, \dots, m-1\}$ then with probability at least $1-\delta$ over the random training set $\mathcal{Z}_m$ any $h \in \mathcal{H}$ satisfies:*

$$\operatorname{err}_u(h) \leq \operatorname{err}_m(h) + \mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m)\right] + \sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \tag{8}$$

Moreover, with probability at least $1-\delta$ any $h \in \mathcal{H}$ satisfies:

$$\operatorname{err}_u(h) \leq \operatorname{err}_m(h) + \hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m) + 2\sqrt{\frac{2N\log(2/\delta)}{(N-1/2)^2}}. \tag{9}$$

Proof. The proof can be found in Sect. 5.4.

We conclude by comparing risk bounds of Theorems 5 and 4:

1. First of all, the upper bound of (9) is computable. This bound is based on the concentration argument, which shows that the expected PRC (appearing in (8)) can be nicely estimated using the training set. Meanwhile, the upper bound of (7) depends on the *unknown* labels of the test set through TRC. In order to make it computable the authors of [10] resorted to the contraction inequality, which allows to drop any dependence on the labels for Lipschitz losses, which is known to be a loose step [17].

2. Moreover, we would like to note that for binary loss function TRC (as well as the Rademacher complexity) does not depend on the labels at all. Indeed, this can be shown by writing $\ell_{01}(y, y') = (1 - yy')/2$ for $y, y' \in \{-1, +1\}$ and noting that σ_i and $\sigma_i y$ are identically distributed for σ_i used in Definition 2. This is not true for PRC, which is *sensitive* to the labels even in this setting. As a future work we hope to use this fact for analysis in the low noise setting [4].

3. The slack term appearing in (8) is significantly smaller than the one of (7). For instance, if $\delta = 0.01$ then the latter is 13 times larger. This is caused by the additive term in symmetrization inequality (2). At the same time, Corollary 1 shows that the complexity term appearing in (8) is at most two times larger than TRC, appearing in (7).

4. Comparison result of Theorem 3 shows that the upper bound of (9) is also tighter than the one which can be obtained using (1) and conditional Rademacher complexity.

5. Similar upper bounds (up to extra factor of 2) also hold for the *excess risk* $\operatorname{err}_u(h_m) - \inf_{h\in\mathcal{H}} \operatorname{err}_u(h)$, where h_m minimizes the training risk err_m over $\mathcal{H}$. This can be proved using a similar argument to Theorem 5.

6. Finally, one more application of the concentration argument can simplify the computation of PRC, by estimating the expected value appearing in Definition 3 with only one random partition of $\mathcal{Z}_m$.

5 Full Proofs

5.1 Proof of Theorem 2

Lemma 3. *For $0 < m \leq N$ let $\mathcal{S}_m := \{s_1, \ldots, s_m\}$ be sampled uniformly without replacement from a finite set of real numbers $\mathcal{C} = \{c_1, \ldots, c_N\} \subset \mathbb{R}$. Then:*

$$\mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\frac{1}{m}\sum_{i=1}^{m} s_i\right] = \frac{1}{\binom{N}{m}}\sum_{\mathcal{S}_m \subseteq \mathcal{C}} \frac{1}{m}\sum_{z \in \mathcal{S}_m} z = \frac{1}{m\binom{N}{m}}\sum_{i=1}^{N}\binom{N-1}{m-1}c_i = \frac{1}{N}\sum_{i=1}^{N} c_i.$$

Proof (of Theorem 2). Fix any *positive* integers n and k such that $n + k = m$, which implies $n < m$ and $k < m = u$. Note that Lemma 3 implies:

$$\bar{f}(\mathcal{Z}_u) = \mathop{\mathbb{E}}_{\mathcal{S}_k}\left[\bar{f}(\mathcal{S}_k)\right], \quad \bar{f}(\mathcal{Z}_m) = \mathop{\mathbb{E}}_{\mathcal{S}_n}\left[\bar{f}(\mathcal{S}_n)\right],$$

where $\mathcal{S}_k$ and $\mathcal{S}_n$ are sampled uniformly without replacement from $\mathcal{Z}_u$ and $\mathcal{Z}_m$ respectively. Using Jensen's inequality we get:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{Z}_m}\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right) &= \mathop{\mathbb{E}}_{\mathcal{Z}_m}\sup_{f\in F}\left(\mathop{\mathbb{E}}_{\mathcal{S}_k}\left[\bar{f}(\mathcal{S}_k)\right] - \mathop{\mathbb{E}}_{\mathcal{S}_n}\left[\bar{f}(\mathcal{S}_n)\right]\right)\\ &\leq \mathop{\mathbb{E}}_{(\mathcal{Z}_m,\mathcal{S}_k,\mathcal{S}_n)}\sup_{f\in F}\left(\bar{f}(\mathcal{S}_k) - \bar{f}(\mathcal{S}_n)\right). \qquad (10)\end{aligned}$$

The marginal distribution of $(\mathcal{S}_k, \mathcal{S}_n)$, appearing in (10), can be equivalently described by first sampling $\mathcal{Z}_m$ from $\mathcal{Z}_N$, then $\mathcal{S}_n$ from $\mathcal{Z}_m$ (both times uniformly without replacement), and setting $\mathcal{S}_k := \mathcal{Z}_m \setminus \mathcal{S}_n$ (recall that $n + k = m$). Thus

$$\mathop{\mathbb{E}}_{(\mathcal{Z}_m,\mathcal{S}_k,\mathcal{S}_n)}\sup_{f\in F}\left(\bar{f}(\mathcal{S}_k) - \bar{f}(\mathcal{S}_n)\right) = \mathop{\mathbb{E}}_{\mathcal{Z}_m}\left[\mathop{\mathbb{E}}_{\mathcal{S}_n}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_m \setminus \mathcal{S}_n) - \bar{f}(\mathcal{S}_n)\right)\middle|\mathcal{Z}_m\right]\right],$$

which completes the proof of the upper bound.

We have shown that for $n \in \{1, \ldots, m-1\}$ and $k := m - n$:

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m}\left[\hat{Q}_{m,n}(F, \mathcal{Z}_m)\right] = \mathop{\mathbb{E}}_{(\mathcal{Z}_k,\mathcal{Z}_n)}\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n)\right), \qquad (11)$$

where $\mathcal{Z}_n$ and $\mathcal{Z}_k$ are sampled uniformly without replacement from $\mathcal{Z}_N$ and $\mathcal{Z}_N \setminus \mathcal{Z}_n$ respectively. Let $\mathcal{Z}_{m-n}$ be sampled uniformly without replacement from $\mathcal{Z}_N \setminus (\mathcal{Z}_n \cup \mathcal{Z}_k)$ and let $\mathcal{Z}_{u-k}$ be the remaining $u-k$ elements of $\mathcal{Z}_N$. Using Lemma 3 once again we get:

$$\mathbb{E}\left[\bar{f}(\mathcal{Z}_{m-n})\middle|(\mathcal{Z}_n, \mathcal{Z}_k)\right] = \mathbb{E}\left[\bar{f}(\mathcal{Z}_{u-k})\middle|(\mathcal{Z}_n, \mathcal{Z}_k)\right].$$

We can rewrite the r.h.s. of (11) as:

$$\begin{aligned}&\mathop{\mathbb{E}}_{(\mathcal{Z}_n,\mathcal{Z}_k)}\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n) + \mathbb{E}\left[\bar{f}(\mathcal{Z}_{u-k}) - \bar{f}(\mathcal{Z}_{m-n})\middle|(\mathcal{Z}_n, \mathcal{Z}_k)\right]\right)\\ &\leq \mathbb{E}\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n) + \bar{f}(\mathcal{Z}_{u-k}) - \bar{f}(\mathcal{Z}_{m-n})\right),\end{aligned}$$

where we have used Jensen's inequality. If we take $n^* = k^* = m/2$ we get

$$\mathop{\mathbb{E}}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F,\mathcal{Z}_m)\right] \le \mathbb{E}\sup_{f\in F}\left(2\bar{f}(\mathcal{Z}_{k^*}\cup\mathcal{Z}_{u-k^*}) - 2\bar{f}(\mathcal{Z}_{n^*}\cup\mathcal{Z}_{m-n^*})\right).$$

It is left to notice that the random subsets $\mathcal{Z}_{k^*}\cup\mathcal{Z}_{u-k^*}$ and $\mathcal{Z}_{n^*}\cup\mathcal{Z}_{m-n^*}$ have the same distributions as $\mathcal{Z}_u$ and $\mathcal{Z}_m$.

5.2 Proof of Theorem 3

Let $m = 2\cdot n$, $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^m$ be i.i.d. Rademacher signs, and $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^m$ be a uniform random permutation of a set containing n plus and n minus signs. The proof of Theorem 3 is based on the coupling of random variables $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$, which is described in Lemma 4. We will need a number of definitions. Consider binary cube $B_m := \{-1,+1\}^m$. Denote $S_m := \{v\in B_m\colon \sum_{i=1}^m v_i = 0\}$, which is a set of all the vectors in B_m having equal number of plus and minus signs. For any $v\in B_m$ denote $\|v\|_1 = \sum_{i=1}^m |v_i|$ and consider the following set:

$$T(v) = \arg\min_{v'\in S_m}\|v-v'\|_1,$$

which consists of the points in S_m closest to v in Hamming metric. For any $v\in B_m$ let $t(v)$ be a random element of $T(v)$, distributed uniformly. We will use $t_i(v)$ to denote i-th coordinate of the vector $t(v)$.

Remark 1. If $v\in S_m$ then $T(v) = \{v\}$. Otherwise, $T(v)$ will clearly contain more than one element of S_m. Namely, it can be shown, that if for some positive integer q it holds that $\sum_{i=1}^m v_i = q$, then q is necessarily even and $T(v)$ consists of all the vectors in S_m which can be obtained by replacing $q/2$ of $+1$ signs in v with -1 signs, and thus in this case $\operatorname{card}\big(T(v)\big) = \binom{(m+q)/2}{q/2}$.

Lemma 4 (Coupling). *Assume that $m = 2\cdot n$. Then the random sequence $t(\boldsymbol{\epsilon})$ has the same distribution as $\boldsymbol{\eta}$.*

Proof. Note that the support of $t(\boldsymbol{\epsilon})$ is equal to S_m. From symmetry it is easy to conclude that the distribution of $t(\boldsymbol{\epsilon})$ is exchangable. This means that it is invariant under permutations and as a consequence uniform on S_m.

Next result is in the core of the multiplicative upper bound (3).

Lemma 5. *Assume that $m = 2\cdot n$. For any $q\in\{1,\dots,m\}$ the following holds:*

$$\mathbb{E}[\epsilon_q | t(\boldsymbol{\epsilon})] = \left(1 - 2^{-m}\binom{m}{n}\right) t_q(\boldsymbol{\epsilon}) \ge \left(1 - 2(2\pi m)^{-1/2}\right) t_q(\boldsymbol{\epsilon}).$$

Proof. We will first upper bound $\mathbb{P}\{\epsilon_q \ne t_q(\boldsymbol{\epsilon}) | t(\boldsymbol{\epsilon}) = \boldsymbol{e}\}$, where $\boldsymbol{e} = \{e_i\}_{i=1}^m$ is (w.l.o.g.) a sequence of n plus signs followed by a sequence of n minus signs.

$$\begin{aligned}\mathbb{P}\{\epsilon_q \ne t_q(\boldsymbol{\epsilon}) | t(\boldsymbol{\epsilon}) = \boldsymbol{e}\} &= \frac{\mathbb{P}\{\epsilon_q \ne t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e}\}}{\mathbb{P}\{t(\boldsymbol{\epsilon}) = \boldsymbol{e}\}} \\ &= \binom{m}{n} 2^{-m} \sum_{\boldsymbol{s}} \mathbb{P}\{\epsilon_q \ne t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e} | \boldsymbol{\epsilon} = \boldsymbol{s}\}, \quad (12)\end{aligned}$$

where we have used Lemma 4 and the sum is over all different sequences of m signs $\boldsymbol{s} = \{s_i\}_{i=1}^m$. For any $\boldsymbol{s}$ denote $S(\boldsymbol{s}) = \sum_{j=1}^n s_j$ and consider terms in (12) corresponding to $\boldsymbol{s}$ with $S(\boldsymbol{s}) = 0$, $S(\boldsymbol{s}) > 0$, and $S(\boldsymbol{s}) < 0$:

Case 1: $S(\boldsymbol{s}) = 0$. These terms will be zero, since $t(\boldsymbol{s}) = \boldsymbol{s}$.

Case 2: $S(\boldsymbol{s}) > 0$. This means that $\boldsymbol{s}$ "has more plus signs than it should" and according to Remark 1 the mapping $t(\cdot)$ will replace several of "+1" with "-1". In particular, if $s_q = -1$ then $t_q(\boldsymbol{s}) = s_q$ and thus the corresponding terms will be zero. If $s_q = 1$ and in the same time $e_q = 1$ the event $\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e}\}$ also can not hold. Moreover, note that identity $\boldsymbol{e} = t(\boldsymbol{s})$ can hold only if $\boldsymbol{e} \in T(\boldsymbol{s})$, which necessarily leads to

$$\big\{j \in \{1, \dots, m\} \colon s_j = -1\big\} \subseteq \big\{j \in \{1, \dots, m\} \colon e_j = -1\big\}. \tag{13}$$

From this we conclude that if $q \in \{1, \dots, n\}$ then all the terms corresponding to $\boldsymbol{s}$ with $S(\boldsymbol{s}) > 0$ are zero. We will use $U_q(\boldsymbol{e})$ to denote the subset of B_m consisting of sequences $\boldsymbol{s}$, such that (a) $S(\boldsymbol{s}) > 0$, (b) $s_q = 1$, and (c) condition (13) holds. It can be seen that if $\boldsymbol{s} \in U_q(\boldsymbol{e})$ then:

$$\mathbb{P}\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e} | \boldsymbol{\epsilon} = \boldsymbol{s}\} = \binom{n + S(\boldsymbol{s})/2}{S(\boldsymbol{s})/2}^{-1}.$$

This holds since, according to Remark 1, $t(\boldsymbol{\epsilon})$ can take exactly $\binom{n+S(\boldsymbol{s})/2}{S(\boldsymbol{s})/2}$ different values, while only one of them is equal to $\boldsymbol{e}$.

Let us compute the cardinality of $U_q(\boldsymbol{e})$ for $q \in \{n+1, \dots, m\}$. It is easy to check that condition $S(\boldsymbol{s}) = 2j$ for some positive integer j implies that $\boldsymbol{s}$ has exactly $n - j$ minus signs. Considering the fact that $s_q = 1$ for $\boldsymbol{s} \in U_q(\boldsymbol{e})$ we have:

$$\operatorname{card}\big(U_q(\boldsymbol{e})\big) = \binom{n-1}{n-j}.$$

Combining everything together we have:

$$\sum_{\boldsymbol{s}\colon\, S(\boldsymbol{s})>0} \mathbb{P}\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e} | \boldsymbol{\epsilon} = \boldsymbol{s}\} = \mathbb{1}\{q > n\} \sum_{j=1}^{n} \frac{\binom{n-1}{n-j}}{\binom{n+j}{j}}.$$

Finally, it is easy to show using induction that:

$$\sum_{j=1}^{n} \frac{\binom{n-1}{n-j}}{\binom{n+j}{j}} = \frac{1}{2}.$$

Case 3: $S(\boldsymbol{s}) < 0$. We can repeat all the steps of the previous case and get:

$$\sum_{\boldsymbol{s}\colon\, S(\boldsymbol{s})<0} \mathbb{P}\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) \cap t(\boldsymbol{\epsilon}) = \boldsymbol{e} | \boldsymbol{\epsilon} = \boldsymbol{s}\} = \frac{1}{2}\mathbb{1}\{q \leq n\}.$$

Accounting for these three cases in (12) we conclude that

$$\mathbb{P}\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) | t(\boldsymbol{\epsilon}) = \boldsymbol{e}\} = \frac{1}{2}\binom{m}{n} 2^{-m} \leq \frac{1}{\sqrt{2\pi m}},$$

where we have used the upper bound on the binomial coefficient from [19, Corollary 2.4]. We can conclude the proof of lemma by writing:

$$\mathbb{E}[\epsilon_q | t(\boldsymbol{\epsilon})] = t_q(\boldsymbol{\epsilon})\left(1 - 2\mathbb{P}\{\epsilon_q \neq t_q(\boldsymbol{\epsilon}) | t(\boldsymbol{\epsilon})\}\right) \geq t_q(\boldsymbol{\epsilon})\left(1 - 2(2\pi m)^{-1/2}\right).$$

Proof (of Theorem 3). First we prove (3). Let $\mathcal{Z}_m = \{z_1, \ldots, z_m\}$. We can write:

$$\hat{Q}_{m,n}(F) = \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\boldsymbol{\epsilon}) f(z_i)\right] \tag{14}$$

$$\leq \left(1 - 2(2\pi m)^{-1/2}\right)^{-1} \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \mathbb{E}[\epsilon_i | t(\boldsymbol{\epsilon})] f(z_i)\right] \tag{15}$$

$$\leq \left(1 + \frac{2}{\sqrt{2\pi m} - 2}\right) \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right], \tag{16}$$

where we have used coupling Lemma 4 in (14), Lemma 5 in (15), and Jensen's inequality in (16). This completes the proof of (3).

Next we prove (4). We have:

$$\left|\hat{Q}_{m,n}(F) - \hat{R}_m(F)\right| = \left|\mathop{\mathbb{E}}_{\boldsymbol{\eta}}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \eta_i f(z_i)\right] - \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right]\right|.$$

Using Lemma 4 and Jensen's inequality we further get:

$$\begin{aligned}
&\left|\hat{Q}_{m,n}(F) - \hat{R}_m(F)\right| \\
&= \left|\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\mathop{\mathbb{E}}_{t}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\boldsymbol{\epsilon}) f(z_i)\middle|\boldsymbol{\epsilon}\right]\right] - \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right]\right| \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\mathop{\mathbb{E}}_{t}\left[\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\boldsymbol{\epsilon}) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right| \middle|\boldsymbol{\epsilon}\right]\right],
\end{aligned} \tag{17}$$

where we have, perhaps misleadingly, denoted the conditional expectation with respect to the uniform choice from $T(\boldsymbol{\epsilon})$ given $\boldsymbol{\epsilon}$ using $\mathbb{E}_t[\,\cdot\,|\boldsymbol{\epsilon}]$. Next we have:

$$\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\boldsymbol{\epsilon}) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right| \leq \left|\sup_{f\in F} \frac{4}{m}\sum_{i\in S(\boldsymbol{\epsilon},t)} \epsilon_i f(z_i)\right|, \tag{18}$$

where $S(\boldsymbol{\epsilon}, t) \subseteq \{1, \ldots, m\}$ is a subset of indices, s.t. $\big(t(\boldsymbol{\epsilon})\big)_i \neq \epsilon_i$ iff $i \in S(\boldsymbol{\epsilon}, t)$. We can continue by writing

$$\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\boldsymbol{\epsilon}) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \epsilon_i f(z_i)\right| \leq \frac{4}{m}\sup_{f\in F}\sum_{i\in S(\boldsymbol{\epsilon},t)} |f(z_i)|. \tag{19}$$

Note that since functions in F are absolutely bounded by B:

$$\sup_{f\in F}\sum_{i\in S(\boldsymbol{\epsilon},t)}|f(z_i)|\le B\cdot\operatorname{card}\left(S(\boldsymbol{\epsilon},t)\right).$$

Returning to (17) and using Remark 1 we obtain:

$$\left|\hat{Q}_{m,n}(F)-2\hat{R}_m(F)\right|\le\frac{4B}{m}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\mathop{\mathbb{E}}_{t}\left[\operatorname{card}\left(S(\boldsymbol{\epsilon},t)\right)|\boldsymbol{\epsilon}\right]\right]=\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\frac{1}{2}\left|\sum_{i=1}^{m}\epsilon_i\right|\right].$$

Khinchin's inequality [15, Lemma 4.1] together with the best known constant due to [12] gives $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left|\sum_{i=1}^{m}\epsilon_i\right|\right]\le\sqrt{m}$, which completes the proof of (4).

5.3 Proof of Lemma 5

Proof. Let $\mathcal{Z}_m=\{z_1,\ldots,z_m\}$. Take F'_m to be a set of two constant functions, $f_1(z)=1$ and $f_2(z)=0$ for all $z\in\mathcal{Z}$. Clearly, $\hat{Q}_{m,n}(F'_m)=0$. In the same time:

$$\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\sup_{f\in F'_m}\frac{2}{m}\sum_{i=1}^{m}\epsilon_i f(z_i)\right]=\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\max\left\{0,\frac{2}{m}\sum_{i=1}^{m}\epsilon_i\right\}\right]\le\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\left|\frac{2}{m}\sum_{i=1}^{m}\epsilon_i\right|\right]\le\frac{2}{\sqrt{m}},$$

where we used Khinchin's inequality. Finally, Khinchin's inequality also gives:

$$\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\max\left\{0,\frac{2}{m}\sum_{i=1}^{m}\epsilon_i\right\}\right]=\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\left|\frac{2}{m}\sum_{i=1}^{m}\epsilon_i\right|\right]\ge\frac{1}{\sqrt{2m}}.$$

Next, let F''_m contain $\binom{m}{m/2}$ functions, such that their projections on $\mathcal{Z}_m$ recover all the permutations of binary vector containing equal number of 0 and 1. Clearly, in this case $\hat{Q}_{m,n}(F''_m)=1$. Straightforward calculations show that in the same time $\hat{R}_m(F''_m)=1-2^{-m}\binom{m}{n}$ and we conclude the proof using upper and lower bounds on the binomial coefficient from [19, Corollary 2.4].

5.4 Proof of Theorem 5

The following version of McDiarmid's bounded difference inequality for the setting of sampling without replacement was presented in [10, Lemma 2] and further improved in [8, Theorem 5]:

Theorem 6 ([10], [8]). *Let $\mathcal{Z}_m$ be sampled uniformly without replacement from a fixed set $\mathcal{Z}_{m+u}\subseteq\mathcal{Z}$ of $m+u$ elements. Let $g\colon\mathcal{Z}^m\to\mathbb{R}$ be a symmetric function s.t. for all $i=1,\ldots,m$ and for all $z_1,\ldots,z_m\in\mathcal{Z}$ and $z'_1,\ldots,z'_m\in\mathcal{Z}$,*

$$\left|g(z_1,\ldots,z_m)-g(z_1,\ldots,z_{i-1},z'_i,z_{i+1},\ldots,z_m)\right|\le c. \tag{20}$$

Then if $m=u$ with probability not less than $1-\delta$ the following holds:

$$g\le\mathbb{E}[g]+\sqrt{\frac{c^2N^3\log(1/\delta)}{8(N-1/2)^2}}.$$

Note that function $\sup_{h\in\mathcal{H}}(\mathrm{err}_h(\mathcal{Z}_u)-\mathrm{err}_h(\mathcal{Z}_m))$ maps $(\mathcal{X}\times\mathcal{Y})^m$ to $\mathbb{R}$ and is of course symmetric. Straightforward calculations show that this function satisfies bounded difference condition (20) with $c=\frac{1}{m}+\frac{1}{u}$ ([10, Inequality 9]). Theorem 6 states that with probability not less than $1-\delta$:

$$\sup_{h\in\mathcal{H}}\left(\mathrm{err}_u(h)-\mathrm{err}_m(h)\right)\le \mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\sup_{h\in\mathcal{H}}\left(\mathrm{err}_u(h)-\mathrm{err}_m(h)\right)\right]+\sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \quad (21)$$

Using upper bound of Theorem 2 with $L_{\mathcal{H}}$ in place of F we complete the proof of (8). Next, consider a symmetric function $-\hat{Q}_{m,n}(L_{\mathcal{H}},\mathcal{Z}_m)$ which also maps $(\mathcal{X}\times\mathcal{Y})^m$ to $\mathbb{R}$. It can be shown again that it satisfies bounded difference condition (20) with $c=\frac{2}{m}$. And thus, Theorem 6 gives that with probability not less than $1-\delta$:

$$\mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\hat{Q}_{m,n}(L_{\mathcal{H}},\mathcal{Z}_m)\right]\le\hat{Q}_{m,n}(L_{\mathcal{H}},\mathcal{Z}_m)+\sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \quad (22)$$

Using this inequality together with (8) in a union bound we obtain the second inequality of the theorem.

Appendix: Improving Lemma 3 of [2]

Let μ be a probability distribution on $\mathcal{Z}$ and $\mathcal{X}_m:=\{X_1,\dots,X_m\}$ be i.i.d. samples selected according to μ. Maximal discrepancy of F was defined in [2] as:

$$\hat{D}_m(F,\mathcal{X}_m)=\sup_{f\in F}\left(\frac{2}{m}\sum_{i=1}^{m/2}f(X_i)-\frac{2}{m}\sum_{i=m/2+1}^{m}f(X_i)\right).$$

It was shown in [2] that if functions in F are uniformly bounded by 1 then:

$$\frac{1}{2}\,\mathbb{E}\left[\hat{R}_m(F,\mathcal{X}_m)\right]-2\sqrt{\frac{2}{m}}\le\mathbb{E}\left[\hat{D}_m(F,\mathcal{X}_m)\right]\le\mathbb{E}\left[\hat{R}_m(F,\mathcal{X}_m)\right]+4\sqrt{\frac{2}{m}}. \quad (23)$$

Since elements in $\mathcal{X}_m$ are i.i.d. the distribution of $\hat{D}_m$ is invariant under their permutations and thus $\mathbb{E}\left[\hat{D}_m(F,\mathcal{X}_m)\right]=\mathbb{E}\left[\hat{Q}_{m,m/2}(F,\mathcal{X}_m)\right]$. Now we can use Theorem 3 to significantly improve bounds in (23):

$$\mathbb{E}\left[\hat{R}_m(F,\mathcal{X}_m)\right]-\frac{2}{\sqrt{m}}\le\mathbb{E}\left[\hat{D}_m(F,\mathcal{X}_m)\right]\le\left(1+\frac{2}{\sqrt{2\pi m}-2}\right)\mathbb{E}\left[\hat{R}_m(F,\mathcal{X}_m)\right].$$

Acknowledgments

The authors are thankful to Marius Kloft and Ruth Urner for useful discussions and to the anonymous reviewers for their comments. GB aknowledges support of the DFG through the FOR-1735 grant. NZ was supported solely by the Russian Science Foundation grant (project 14-50-00150).

References

1. Bartlett, P., Bousquet, O., Mendelson, S.: Local rademacher complexities. The Annals of Statistics, 33(4), 1497–1537 (2005)
2. Bartlett, P., Mendelson, S.: Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3, 463–482 (2001)
3. Blum, A., Langford, J.: PAC-MDL Bounds. In: COLT 2003, pp. 344–357 (2003)
4. Boucheron, S., Lugosi, G., Bousquet, O.: Theory of classification: a survey of recent advances. ESAIM: Probability and Statistics, 9, 323–375 (2005)
5. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press (2013)
6. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press (2006)
7. Cortes, C., Mohri, M.: On transductive regression. In: NIPS 2006, 305–312 (2007)
8. Cortes, C., Mohri, M., Pechyony, D., Rastogi, A.: Stability analysis and learning bounds for transductive regression algorithms. CoRR **abs/0904.0814** (2009)
9. Derbeko, P., El-Yaniv, R., Meir, R.: Explicit learning curves for transduction and application to clustering and compression algorithms. Journal of Artificial Intelligence Research, 22(1), 117–142 (2004)
10. El-Yaniv, R., Pechyony, D.: Transductive rademacher complexity and its applications. Journal of Artificial Intelligence Research, 35(1), 193–234 (2009)
11. Gross, D., Nesme, V.: Note on sampling without replacing from a finite collection of matrices. `http://arxiv.org/abs/1001.2738v2` (2010)
12. Haagerup, U.: The best constants in Khinchine inequality. Studia Mathematica, 70(3), 231–283 (1981)
13. Koltchinskii, V.: Oracle inequalities in empirical risk minimization and sparse recovery problems. Springer (2011)
14. Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In: Gine. D.E., Wellner, J. (eds.) High Dimensional Probability, II, pp. 443–457. Birkhauser (1999)
15. Ledoux, M., Talagrand, M.: Probability in Banach Space. Springer-Verlag (1991)
16. Magdon-Ismail, M.: Permutation complexity bound on out-sample error. In: Advances in Neural Information Processing Systems (NIPS 2010), pp. 1531–1539 (2010)
17. Mendelson, S.: Learning without Concentration. CoRR **abs/1401.0304** (2014)
18. Pechyony, D.: Theory and Practice of Transductive Learning. PhD thesis (2008)
19. Stanica, P.: Good lower and upper bounds on binomial coefficients. Journal of Inequalities in Pure and Applied Mathematics, 2(3) (2001)
20. Tolstikhin, I., Blanchard, G., Kloft, M.: Localized complexities for transductive learning. In: COLT 2014, pp. 857–884 (2014)
21. Van der Vaart, A. W., Wellner, J.: Weak Convergence and Empirical Processes: With Applications to Statistics. Springer (2000)
22. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons (1998)